Biomedical word sense disambiguation with word embeddings

Rui Antunes and Sérgio Matos

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
ruiantunes@ua.pt, aleixomatos@ua.pt

Abstract. There is a growing need for automatic extraction of information and knowledge from the increasing amount of biomedical and clinical data produced, namely in textual form. Natural language processing comes in this direction, helping in tasks such as information extraction and information retrieval. Word sense disambiguation is an important part of this process, being responsible for assigning the proper concept to an ambiguous term.

In this paper, we present results from machine learning and knowledge-based algorithms applied to biomedical word sense disambiguation. For the supervised machine learning algorithms we used word embeddings, calculated from the full MEDLINE literature database, as global features and compare the results to the use of local unigram and bigram features.

For the knowledge-based method we represented the textual definitions of biomedical concepts from the UMLS database as word embedding vectors, and combined this with concept associations derived from the MeSH term co-occurrences.

Both the machine learning and the knowledge-based results indicate that word embeddings are informative and improve the biomedical word disambiguation accuracy. Applied to the reference MSH WSD data set, our knowledge-based approach achieves 85.1% disambiguation accuracy, which is higher than some previously proposed approaches that do not use machine-learning strategies.

Keywords: biomedical word sense disambiguation, word embeddings

1 Introduction

Large volumes of biomedical data are produced every day, and this is accompanied by an increasing amount of textual data, mostly in the form of scientific publications. In order to efficiently treat and interpret these data it is necessary to create tools that automatically do this job, reducing the human efforts. This led to the application of text mining methods for extracting information from the literature and linking that to repositories of biomedical data [1].

Word Sense Disambiguation (WSD), an important subtask of Natural Language Processing (NLP) [2], is a challenging task that consists of finding the

correct sense of an ambiguous term. Usually, this is achieved using the surrounding context of the term. Currently, there are mainly two distinct approaches for WSD, those based on Machine Learning (ML) algorithms and the ones based on knowledge sources. The ML approaches can follow supervised, semi-supervised or unsupervised algorithms, with supervised classification approaches currently offering the best results, achieving macro and micro accuracy around 96% on the MSH WSD data set using a Support Vector Machine (SVM) classifier [3].

Knowledge-based approaches to WSD have also attracted large interest, as these approaches are usually less dependent on training data, which may lead to better generalization when compared to supervised learning algorithms. The use of multiple knowledge databases brings benefits to the problem of concept disambiguation [4]. WordNet [5] is a large knowledge database of the English language that has been extensively applied for word sense disambiguation [2]. In the case of biomedical texts, the largest and most relevant knowledge database is the Unified Medical Language System (UMLS) [6], which offers a rich integrated metathesaurus and semantic network for the biomedical domain. In this work we used the Medical Subject Headings (MeSH), a hierarchically-organized biomedical vocabulary resource used by the MEDLINE database to index scientific publications, and which is part of the UMLS metathesaurus.

Word embeddings [7] is a recent technique that consists in deriving vector representations of the words within an unlabelled corpus. These vectors can be used for different NLP tasks, namely for the disambiguation process. We used them as global features in the ML classification problem. In our case, these features showed to be almost as effective as local features, such as unigrams and bigrams. Also, we made use of the word embeddings in our knowledge-based approach to represent concepts, and the textual context of ambiguous words, as embedded vectors that can be directly compared. In [3], the authors present a work on supervised biomedical word sense disambiguation applied to the MSH WSD data set, exploring the combination of unigrams as local features and word embeddings as global features. Other approaches using word embeddings for word sense disambiguation have also been proposed by Wu et al. [8], and Taghipour and Ng [9].

In this work, we applied knowledge-based methods and machine learning techniques to the MSH WSD data set in order to measure the WSD accuracies. The UMLS database were used to extract textual definitions of biomedical concepts. Also, we used the co-occurrences of the MeSH descriptors[1] to derive concept-concept associations between. The ML classifiers used in this experiment were the decision tree, the k-nearest neighbours, and the linear SVM with stochastic gradient descent. Textual data from the MEDLINE database were used to generate the word embeddings, which were used in the machine learning and knowledge-based approaches.

[1] `https://ii.nlm.nih.gov/MRCOC.shtml`

2 Methods

2.1 The MSH WSD data set

The MSH WSD data set was automatically generated using the UMLS metathesaurus and MEDLINE citations [10]. The data consist of scientific abstracts, each with one ambiguous term identified and mapped to the correct sense. It contains 203 ambiguous terms with a total of 423 distinct senses. Most terms (189) have only two different meanings, 12 terms have three different meanings, and the remaining 2 terms have four and five different meanings. The dataset contains around 37 thousand abstracts, each representing an ambiguity example for a term, therefore averaging 187 ambiguity examples per ambiguous term.

Since we extracted textual definitions and MeSH relations from the UMLS database, not all concepts of the MSH WSD data set were present. Thus, a minor part containing 12 terms[2] of the MSH WSD data set were not used for this disambiguation task. All the presented results do not include these terms.

2.2 Machine Learning

For each ambiguous term, we applied 5-fold cross-validation to subdivide the corresponding abstracts for training and testing the model. A bag-of-words model was used to represent the texts, with local features acquired from the context, namely unigrams and bigrams, with tf-idf weighting. We also applied supervised ML algorithms using word embedding vectors, calculated from the full MEDLINE, as global features. A list of 364 stopwords obtained from the UMLS repository was used to filter out very frequent words in the corpus. All these tasks were implemented using the framework Scikit-learn [11], a machine-learning library for the Python programming language. Word embedding models were obtained with the Word2Vec [7] implementation in the Gensim framework [12].

We tested three machine learning classifiers: decision tree classifier, k-nearest neighbours, and linear SVM with stochastic gradient descent. The local features used were unigrams and bigrams, and the global features used were the word embeddings from the full MEDLINE.

The word embedding models were calculated with PubMed articles, which are specific to biomedical domain, from the full MEDLINE. Around 20 million abstracts corresponding to the years 1900 to 2015 were used, containing around 800 thousand distinct words. We trained six models, with windows of five, twenty and fifty words and for feature vectors of sizes 100 and 300. Each abstract, instance of the MSH WSD data set, was represented by the weighted average of the embedding vectors of the containing words, with the tf-idf value of each word used as weight.

[2] Terms not considered: Ca; CNS; Crown; DBA; FAS; Gamma-Interferon; Hybridization; ITP; PCP; Plaque; Pneumocystis; Semen

2.3 Knowledge-based

We developed a knowledge-based method to choose the most related concepts from a text and which was applied in the disambiguation task. From the UMLS database we extracted all the available concept textual definitions. Additionally, we used the co-occurrence counts of MeSH terms in MEDLINE articles[3] to calculate the normalized Pointwise Mutual Information (nPMI) as an association metric between all pairs of MeSH terms. Since the MSH WSD data set uses UMLS Concept Unique Identifiers (CUIs) to identify the distinct term senses, we used the MeSH to CUI mapping in UMLS to translate these MeSH term associations to (UMLS) concept-concept associations.

We used the same word embedding models as described above for the machine learning approach. Each specific CUI was represented as an embedding vector calculated as the tf-idf weighted average of the words in the concept definition, therefore mapping each concept to an high-dimensional vector. Using the same approach we were able to calculate an embedding vector for each abstract in the MSH WSD data set. Thus, it was possible to infer the most related sense for an ambiguous term by measuring the cosine similarity between its textual context and each possible UMLS concept, selecting the most similar one.

Additionally, we extended this document-concept similarity score using the concept associations obtained from the MeSH co-occurrences, as shown in equation 1.

$$score(CUI) = \frac{1}{N} \sum_{j} nPMI(CUI, CUI_j) \cdot CS(\boldsymbol{t}, \boldsymbol{CUI_j}) \tag{1}$$

According to equation 1, for each possible CUI of an ambiguous target term is assigned a score given by the average of the cosine similarities between the term context vector $\boldsymbol{t}$ and the concept vector of all the concepts $\boldsymbol{CUI_j}$, weighted by the concept association score $nPMI(CUI, CUI_j)$. Each considered CUI has a $nPMI$ value equal to a unit in relation to himself. As before, the concept with highest score is selected as the correct sense for the ambiguous term.

3 Results

Table 1 shows that the state-of-the-art results for this problem can be almost reproduced using simple word-based features. It is also noticeable that bigram features contribute only slightly to the results, and unigram features alone achieve almost as good if not better results than the combination of unigram and bigrams. Also, comparing these results with Table 2, one can observe that word embedding features alone allow obtaining results that are very close to the best results obtained with unigram features.

With the machine learning classifiers the highest accuracy, 94.7%, was obtained with unigram features alone, using the support vector machine linear

[3] `https://ii.nlm.nih.gov/MRCOC.shtml`

classifier. On the other hand, using only global features the accuracies were similar, and the highest accuracy, 94.0%, was also obtained using the support vector machine classifier.

Table 1. Accuracies using local features. Results shown are the average across five folds. U: Unigrams; B: Bigrams; DT: Decision Tree; kNN: k-Nearest Neighbour (k=5); SVM: linear Support Vector Machine with stochastic gradient descent.

	U	B	U+B
DT	0.903	0.862	0.901
kNN	0.913	0.918	0.924
SVM	**0.947**	0.931	0.946

Table 2. Accuracies using word embedding models from the full MEDLINE as global features. S: Size; W: Window; DT: Decision Tree; kNN: k-Nearest Neighbour (k=5); SVM: linear Support Vector Machine with stochastic gradient descent.

	S100			S300		
	W5	W20	W50	W5	W20	W50
DT	0.907	0.909	0.909	0.910	0.911	0.909
kNN	0.931	0.933	0.933	0.931	0.931	0.931
SVM	0.931	0.934	0.937	0.934	0.938	**0.940**

In Table 3 the knowledge-based results are presented. One can see that these results are only about 10% below the machine learnings results, since it is a more generalized method that do not use train data from the data set to predict the correct meanings. We applied a threshold to the concept association nPMI score, in order to filter the associated concepts that contribute to the final score for a CUI (see equation 1). A smaller value for the nPMI threshold means that more related concepts contribute to the final score, and the results show that using more related concepts, and not only the ones with stronger association score, improves the disambiguation accuracy. The highest accuracy, 85.1%, was obtained with a nPMI threshold of 0.30 using the word embedding model with a size vector of 100 and a window of 50 words.

Table 3. Accuracies using the CUI definitions, the CUI relations from the UMLS and word embeddings from the full MEDLINE. CS: cosine similarity between term context vector and concept vector only; nPMI $\geq$ *thresh*: cosine similarity plus related concepts with a nPMI value higher than the threshold; S: Size; W: Window; nPMI: normalized Pontwise Mutual Information.

	S100			S300		
	W5	W20	W50	W5	W20	W50
CS	0.800	0.812	0.813	0.799	0.811	0.810
nPMI $\geq$ 0.9	0.799	0.812	0.813	0.799	0.811	0.809
nPMI $\geq$ 0.8	0.799	0.812	0.813	0.799	0.812	0.810
nPMI $\geq$ 0.7	0.797	0.811	0.813	0.798	0.811	0.809
nPMI $\geq$ 0.6	0.783	0.798	0.799	0.785	0.797	0.795
nPMI $\geq$ 0.5	0.789	0.803	0.805	0.790	0.802	0.798
nPMI $\geq$ 0.4	0.816	0.829	0.831	0.817	0.826	0.826
nPMI $\geq$ 0.3	0.835	0.849	**0.851**	0.835	0.846	0.844
nPMI $\geq$ 0.2	0.827	0.842	0.844	0.826	0.838	0.837

4 Conclusions

As has been previously shown, machine learning algorithms outperform the knowledge-based algorithms in biomedical word sense disambiguation. However, the latter have the advantage of being directly applied to any ambiguous term, since they do not rely on training data. Our approach achieves a robust disambiguation performance that is on par with the best methods that do not use annotated data in a supervised setting, and slightly above the results obtained with the Automatic Extracted Corpus (AEC) [10], which can be applied to obtain training data from MEDLINE to create the disambiguation classifiers on-the-fly, therefore reducing the need for pre-compiled training data. Tulkens [13] obtained a disambiguation accuracy of 84% with a knowledge-based method applied to the same data set using word embeddings from BioASQ corpora. In a recent work, Sabbir et al. [14] combined a knowledge-approach with neural concept embeddings and distant supervision, achieving an accuracy of 92%.

One of the limitations of our approach is that not all UMLS concepts have rich definitions. Also, some concepts of the MSH WSD data are not present in the UMLS database, leading to an incapacity of disambiguation. As future work, we will investigate ways of overcoming this by constructing concept vectors from associated MEDLINE texts.

Acknowledgments

This work was supported by Portuguese National Funds through FCT - Foundation for Science and Technology, in the context of the project IF/01694/2013. Sérgio Matos is funded under the FCT Investigator programme.

This is a post-peer-review, pre-copyedit version of an article published in "Advances in Intelligent Systems and Computing book series (AISC, volume 616)". The final authenticated version is available online at: `https://doi.org/10.1007/978-3-319-60816-7_33`.

References

1. Campos, D., Matos, S., Oliveira, J. L.: A modular framework for biomedical concept recognition. BMC Bioinformatics, 14.1:281 (2013)
2. Navigli, R.: Word sense disambiguation: a survey. ACM Computing Surveys, 41.2:10 (2009)
3. Yepes, A. J.: Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. arXiv:1604.02506v3 (2016)
4. Tsai, C. T., Roth, D.: Concept grounding to multiple knowledge bases via indirect supervision. Transactions of the Association for Computational Linguistics, 4:141–154 (2016)
5. Fellbaum, C.: WordNet: an electronic lexical database. Cambridge: MIT Press (1998)
6. Bodenreider, O.: The unified medical language systems (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32.1:267–270 (2004)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 3111:3119 (2013)
8. Wu, Y., Xu, J., Zhang, Y., Xu, H.: Clinical abbreviation disambiguation using neural word embeddings. ACL-IJCNLP, 171–176 (2015)
9. Taghipour, K., Ng, H. T.: Semi-supervised word sense disambiguation using word embeddings in general and specific domains. HLT-NAACL, 314–323 (2015)
10. Yepes, A. J., McInnes, B. T., Aronson, A. R.: Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics, 12.1:223 (2011)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine Learning in Python. Journal of Machine Learning Research, 12:2825–2830 (2011)
12. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50 (2010)
13. Tulkens, S., Šuster, S., Daelemans, W.: Using distributed representations to disambiguate biomedical and clinical concepts. arXiv:1608.05605v1 (2016)
14. Sabbir, A. K. M., Yepes, A. J., Kavuluru, R.: Knowledge-based biomedical word sense disambiguation with neural concept embeddings and distant supervision. arXiv:1610.08557v3 (2017)